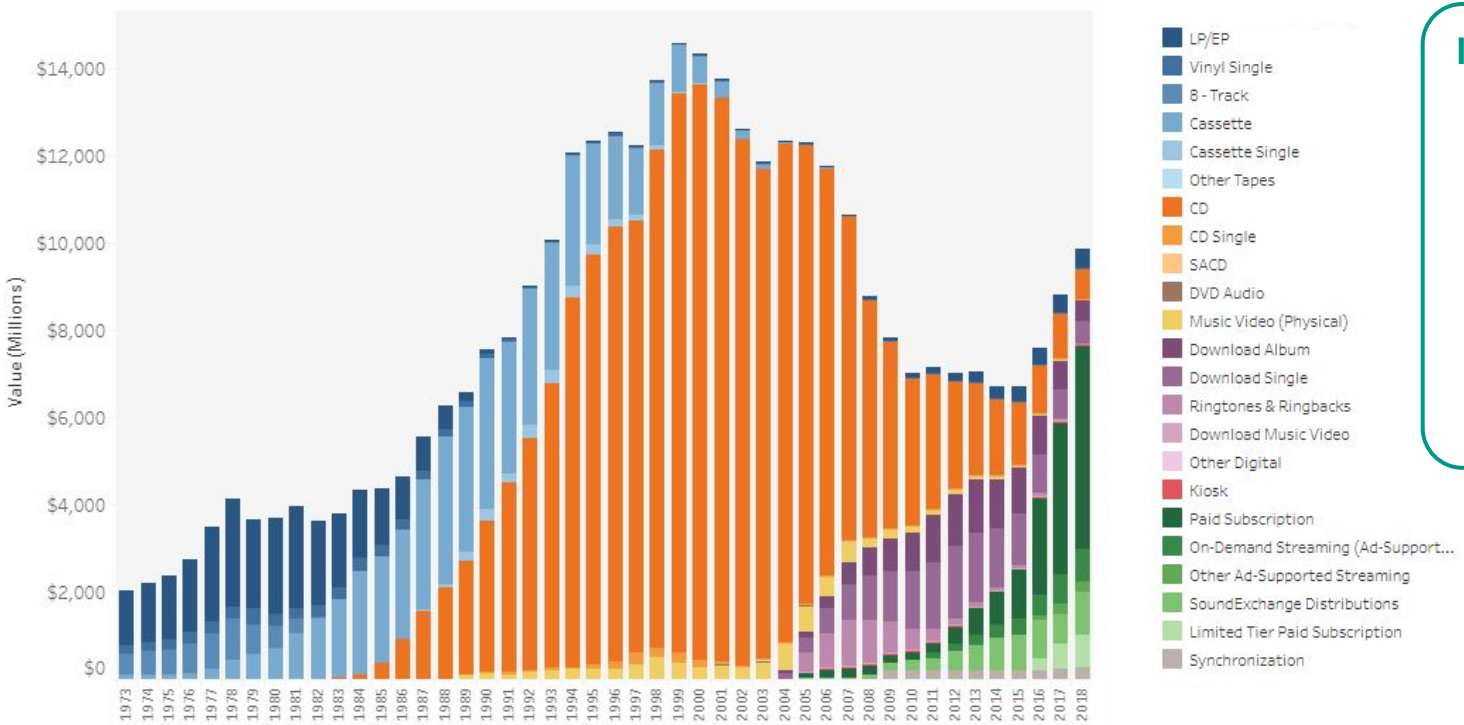
Music Information Retrieval with Neural Nets

W210 Capstone Project, Week 05

Madeleine Bulkow | Kuangwei Huang | Weixing Sun

Background: Music Industry

U.S. Recorded Music Revenues by Format 1973 to 2018



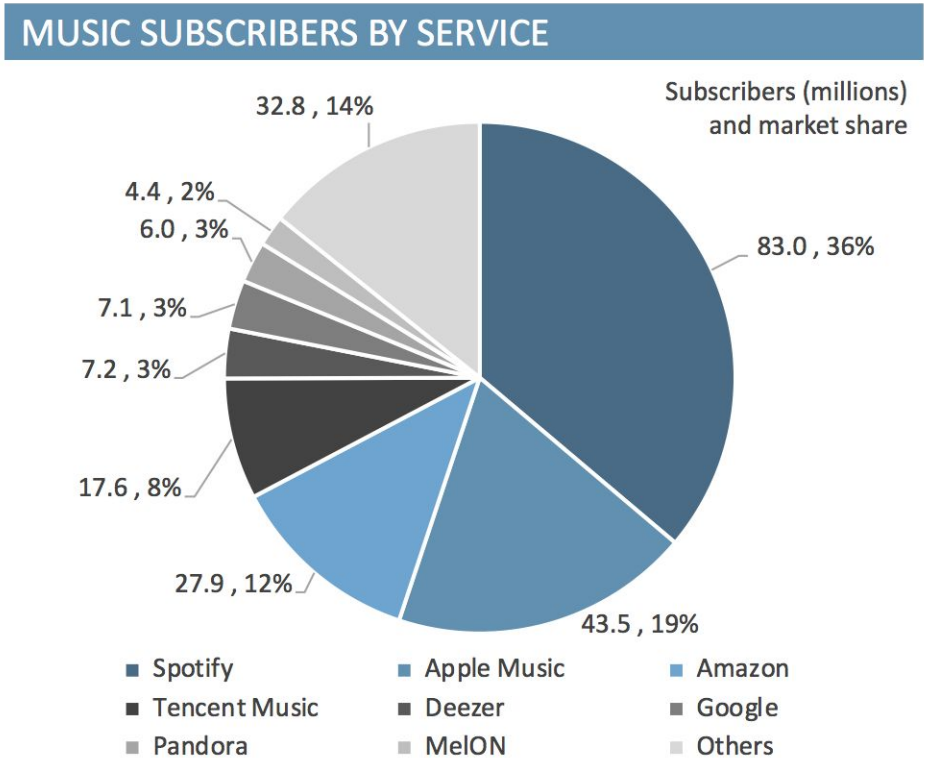
In 2018:

- $9.2 billion
- 12% YoY growth
- 75% streaming
- > 50 mil paid subscriptions

Source: https://www.riaa.com/u-s-sales-database/

Streaming Music: Opportunities

- Globally streaming revenues of $8.9B in 2018
- Streaming revenues grew by 34.0% in 2018
- Platforms compete on personalized content and “discovery”
- Recommender systems, traditionally content-agnostic
- Opportunity for content-based recommendation using deep learning
- Song profiling, akin to NLP word embeddings
- Future: GANs for music generation



Source: https://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2019
Source: https://www.midiaresearch.com/app/uploads/2018/09/midia-mid-year-2018-subscriber-mareket-shares.png

Build Plan

"deep learning techniques

↓

content-based profiling of songs

↓

new approaches to song recommenders"

- Start with genre classification
- Converting Audio samples into Spectrograms
- Convolutional Neural Networks (CNN) for classification
- Song embeddings
- New approaches to recommenders for streaming

Baseline

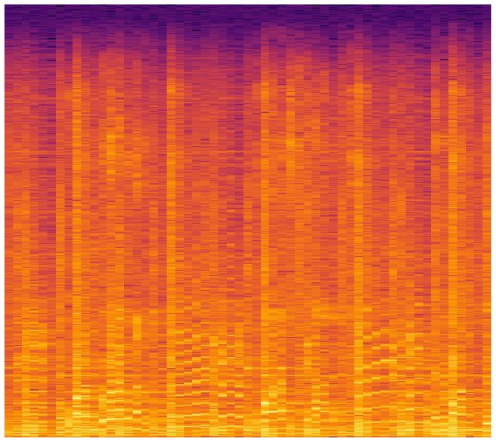
- GTZAN music database: 1000 snippets
- Audio transferred to spectrograms
- Features obtained as inputs
- NN Deep Learning for genre classification: test accuracy at 0.695

Future Plan

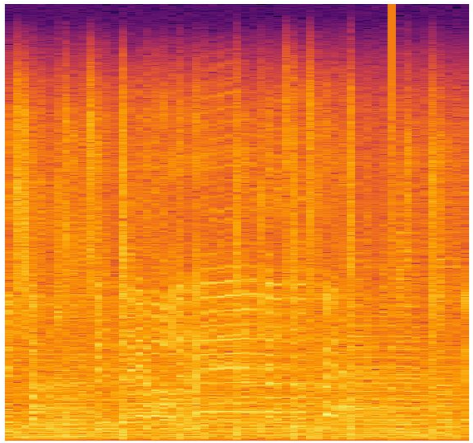
- More database to discover
- Better CNN designs
- RNN and other ML methods
- More functions: Artist and song detections

Mel-Frequency Cepstral Coefficients (MFCCs):
A small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice.

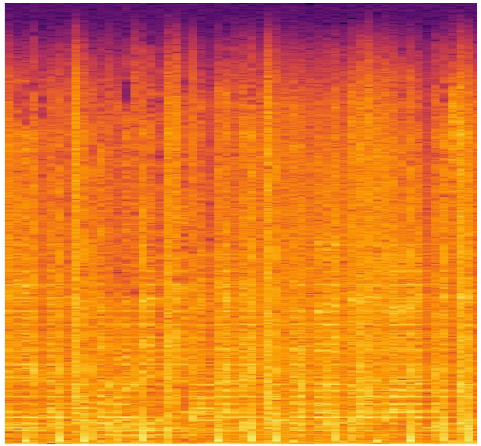
Example: Spectrograms



Blues



Rock



Hiphop

	filename	chroma_stft	rmse	spectral_centroid	spectral_bandwidth	rolloff	zero_crossing_rate	mfcc1	mfcc2	mfcc3	...
0	blues.00043.au	0.399025	0.127311	2155.654923	2372.403604	5012.019693	0.087165	-109.165355	100.621500	-8.614721	...
1	blues.00012.au	0.269320	0.119072	1361.045467	1567.804596	2739.625101	0.069124	-207.208080	132.799175	-15.438986	...
2	blues.00026.au	0.278484	0.076970	1198.607665	1573.308974	2478.376680	0.051988	-284.819504	108.785628	9.131956	...
3	blues.00077.au	0.408876	0.243217	2206.771246	2191.473506	4657.388504	0.111526	-29.010990	104.532914	-30.974207	...
4	blues.00084.au	0.396258	0.235238	2061.150735	2085.159448	4221.149475	0.113397	-38.965941	112.039843	-31.817035	...

Snippet samples: https://drive.google.com/open?id=16jOXcRsmrgrPE54x26R-U-yBA2CHS0DJ

Basic NN Deep Learning

28 attributes 1D layer → Hidden layers → Output: genre → overfitting?

1st accuracy: ~70% → 2nd accuracy (one more hidden layer): ~72%

Building our Network

```
from keras import models
from keras import layers

model = models.Sequential()
model.add(layers.Dense(256, activation='relu', input_shape=(X_train.shape[1],)))

model.add(layers.Dense(128, activation='relu'))

model.add(layers.Dense(64, activation='relu'))

model.add(layers.Dense(10, activation='softmax'))
```

```
Epoch 19/20
800/800 [==============================] - 0s 26us/step - loss: 0.2192 - acc: 0.9600
Epoch 20/20
800/800 [==============================] - 0s 20us/step - loss: 0.2228 - acc: 0.9500
```

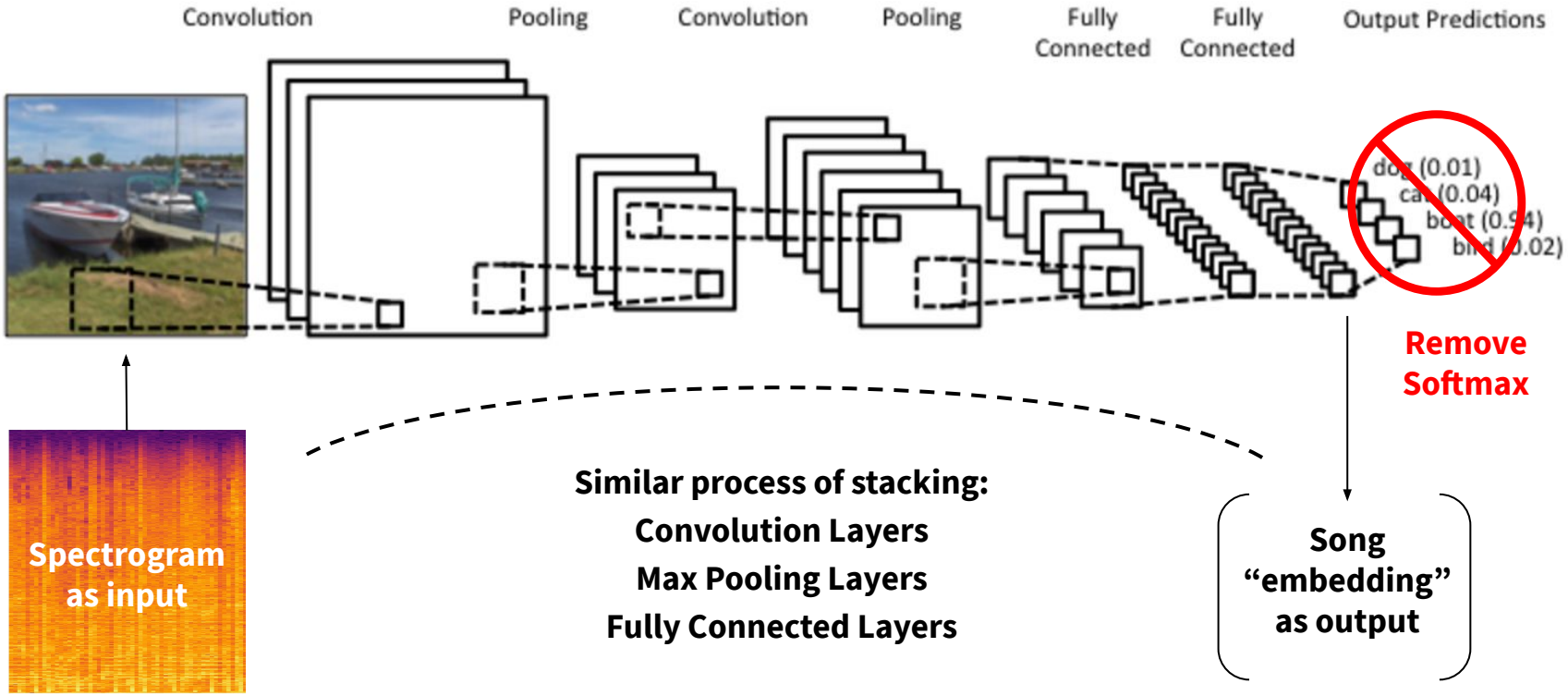
```
test_loss, test_acc = model.evaluate(X_test,y_test)
```

```
200/200 [==============================] - 0s 26us/step
```

```
print('test_acc: ',test_acc)
```

```
test_acc:  0.695
```

CNN: Song Embeddings



Source: http://www.wildml.com/category/neural-networks/convolutional-neural-networks/

https://freemusicarchive.org/genre/Electronic

Data Sources

Source Name	Link	Data Available	Size	Restrictions and Limitations
Spotify	x	Songs in .ogg format, metadata.	35 million songs	Songs are DRM encrypted.
Million Songs Database	x	Metadata, 7 digital song id.	1 million songs	No audio, 7 digital link requires API key.
GTZAN	x	Songs in .wav format, genre labels.	1000 snippets, 30s each.	Small, no titles or artists.
Free Music Archive (FMA)	x	Songs in mp3 format format, metadata.	100,000 songs	
Youtube	x	Videos containing music.	300 million videos labelled "music".	Need to extract mp3, metadata availability variable.

Next Steps

- Improve overfitting issues from 1st attempt with CNN
- Truncating vs sampling rates vs different ways of "summarizing" the data
- Start on song embeddings

Questions?

W210 Capstone Project, Week 05

Madeleine Bulkow | Kuangwei Huang | Weixing Sun

References

1. McFee, et Al, 2015, “librosa: Audio and Music Signal Analysis in Python”, Proceedings in the 14th Python in Science Conference
2. Graves, et Al, 2006, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks”
3. Wang, “An Industrial-Strength Audio Search Algorithm”

Question1: spectrogram size: (128, 1292), sr=22050Hz, too many data points?

Question2: mfcc parameter numbers?

Remove softmax?